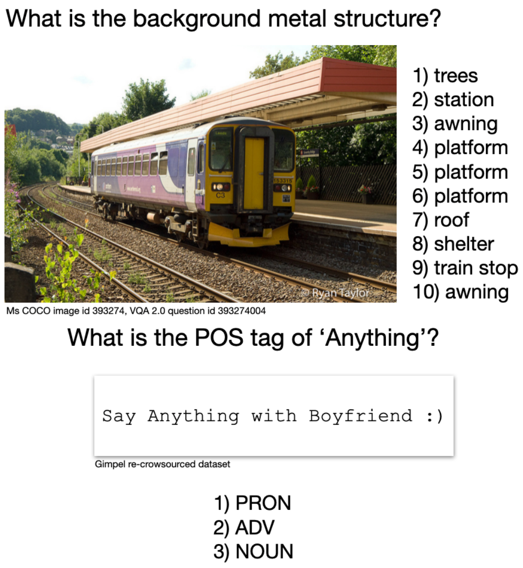
We Need to Consider Disagreement in Evaluation

Valerio Basile[*♣], Michael Fell[♣], Tommaso Fornaciari[♦], Dirk Hovy[♦],
Silviu Paun[♥], Barbara Plank[♠], Massimo Poesio[♥], Alexandra Uma[♥]
[♣]University of Turin, [♦]Bocconi University
[♥]Queen Mary University of London, [♠]IT University of Copenhagen
[♣]{valerio.basile, michaelkurt.fell}@unito.it
[♦]{dirk.hovy, fornaciari.tommaso}@unibocconi.it
[♥]{s.paun, m.poesio, a.n.uma}@qmul.ac.uk, [♠]bplank@itu.dk

Abstract

Evaluation is of paramount importance in data-driven research fields such as Natural Language Processing (NLP) and Computer Vision (CV). But current evaluation practice in NLP, except for end-to-end tasks such as machine translation, spoken dialogue systems, or NLG, largely hinges on the existence of a single "ground truth" against which we can meaningfully compare the prediction of a model. However, this assumption is flawed for two reasons. 1) In many cases, more than one answer is correct. 2) Even where there is a single answer, disagreement among annotators is ubiquitous, making it difficult to decide on a gold standard. We discuss three sources of disagreement: from the annotator, the data, and the context, and show how this affects even seemingly objective tasks. Current methods of adjudication, agreement, and evaluation ought to be reconsidered at the light of this evidence. Some researchers now propose to address this issue by minimizing disagreement, creating cleaner datasets. We argue that such a simplification is likely to result in oversimplified models just as much as it would do for end-to-end tasks such as machine translation. Instead, we suggest that we need to improve today's evaluation practice to better capture such disagreement. Datasets with multiple annotations are becoming more common, as are methods to integrate disagreement into modeling. The logical next step is to extend this to evaluation.

1 Introduction



What is the background metal structure?
1) trees
2) station
3) awning
4) platform
5) platform
6) platform
7) roof
8) shelter
9) train stop
10) awning
© Ryan Taylor
Ms COCO image id 393274, VQA 2.0 question id 393274004
What is the POS tag of 'Anything'?
Say Anything with Boyfriend :)
Gimpel re-crowsourced dataset
1) PRON
2) ADV
3) NOUN

Figure 1: What is the ground truth? Examples from VQA v2 (Goyal et al., 2017) and (Gimpel et al., 2011).

Evaluation is of paramount importance to Natural Language Processing (NLP) and Computer Vision (CV). Automatic evaluation is the primary mechanism to drive and measure progress due to its simplicity and efficiency (Resnik and Lin, 2010; Church and Hestness, 2019). However, today's evaluation practice for virtually all NLP tasks concerned with a fundamental aspect of language interpretation–POS tagging, word sense disambiguation, named entity recognition, coreference, relation extraction, natural language inference, or sentiment analysis– is seriously flawed: the candidate hypotheses of a system (i.e., its predictions) are compared against an evaluation set that is assumed to encode a "ground truth" for the modeling task. Yet this evaluation model is outdated and needs reconsideration. The notion of a single correct answer ignores the subjectivity and complexity of many tasks, and focuses on "easy", low-risk evaluation, holding back progress in the field. We discuss three sources of disagreement: from the annotator, the data, and the context.

The underlying assumption of the current approach is that the evaluation set represents the

[*] Authors in alphabetical order.

Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, pages 15–21
August 5–6, 2021. ©2021 Association for Computational Linguistics

best possible approximation of the truth about a given phenomenon, or at least a reasonable one. This ground truth is usually obtained by developing an annotation scheme for the task aiming to achieve the highest possible agreement between human annotators (Artstein and Poesio, 2008). Disagreements between annotators are either reconciled by hand or aggregated (particularly in the case of crowdsourced annotations) to extract the most likely or agreed-upon choices (Hovy et al., 2013; Passonneau and Carpenter, 2013; Paun et al., 2018). This aggregated data is referred to as "gold standard" (see Ide and Pustejovsky (2017) for an in-depth analysis of annotation methodology).

However, there is plenty of evidence that gold labels are an idealization, and that unreconcilable disagreement is abundant. Figure 1 shows two examples from CV and NLP. This is particularly true for tasks involving highly subjective judgments, such as hate speech detection (Akhtar et al., 2019, 2020) or sentiment analysis (Kenyon-Dean et al., 2018). However, it is not a trivial issue even in more linguistic tasks, such as part-of-speech tagging (Plank et al., 2014), word sense disambiguation (Passonneau et al., 2012; Jurgens, 2013), or coreference resolution (Poesio and Artstein, 2005; Recasens et al., 2011). Systematic disagreement also exists in image classification tasks, where labels may overlap (Rodrigues and Pereira, 2018; Peterson et al., 2019). Disagreement and task difficulty and subjectivity also challenge traditional agreement measures (Artstein and Poesio, 2008). High agreement is typically used as a proxy for data quality. However, it obscures possible sources of disagreement (Poesio and Artstein, 2005). We summarize some of the evidence on disagreement in Section 2.

The need for metrics not based on the assumption that a gold standard exists has long been accepted for end-to-end tasks, particularly those involving an aspect of natural language generation, such as conversational agents, machine translation, surface realisation, image captioning, or summarization. Metrics such as BLEU for machine translation/generation, ROUGE for summarization, or NDCG for ranking Web searches all support more than one gold standard reference. Shared tasks in this areas (particularly on paraphrasing), have also considered the role of disagreement in their evaluation metrics (Butnariu et al., 2009; Hendrickx et al., 2013). Variability in the annotation is a feature of many such tasks (see, e.g., van der Lee et al. (2019) for agreement issues in generated text evaluation) even though many corpora still may come with single references due to data collection costs. High agreement is disfavored, and even bears risks of non-natural, highly homogenized system outputs for generation tasks (Amidei et al., 2018). The main argument of this position paper is that we should recognize that the same issues, if perhaps in less extreme version, apply to the analysis tasks we discuss here.

In recent years, proposals have been put forward to consider the disagreement as informative content that can be leveraged to improve task performance (Plank et al., 2014; Aroyo and Welty, 2015; Jamison and Gurevych, 2015). Uma et al. (2020) and Basile (2020) investigated the impact of disagreement-informed data on the quality of NLP evaluation, and found it to be beneficial and providing complementary information, as further discussed in Section 3. This led them to organize a first shared task on learning from disagreement and providing non-aggregated benchmarks for evaluation (Uma et al., 2021).

In contrast with this trend, Bowman and Dahl (2021) recently proposed to study biases and artifacts in data to eliminate them. Beigman Klebanov and Beigman (2009) adopt a slightly softer stance, proposing to only evaluating on "easy" (as in, highly agreed upon) instances. Based on the evidence about the prevalence of disagreement in NLP judgments, we argue against this approach. First, it leads to information loss in the attempt to reducing noise in the data. Second, it is unnecessary: while evaluation methods that include disagreement are not yet established, several methodologies already do exist. Removing the disagreement might lead to better evaluation scores, but it fundamentally hides the true nature of the task we are trying to solve.

2 Disagreement in NLP

In this section, we outline three possible sources of disagreement. Afterward, we describe how disagreement has been studied in objective and arguably more subjective tasks in NLP.

2.1 Sources of Disagreement

Annotation implies an interaction between the human judge, the instance which has to be evaluated, and the moment/context in which the process takes place. For each instance, the annotation outcome depends on these three elements, assuming the task is properly defined, designed, and carried out, e.g., in terms of quality control. We summarize these potential sources of disagreement as follows:

Individual Differences. World perception is a personal and intrinsically private experience. To some extent, this experience can be traced back to a common ground, but margins of subjectivity remain. These margins are relatively limited when they concern matters of fact, but they snowball when opinions, values, and sentiments come into play. In NLP, many annotation tasks rely on personal opinions and judgment, despite uniform instructions for annotators. For example, in hate speech detection or sentiment analysis, different annotators might have very different perspectives regarding what is hateful or negative, respectively. Individual differences remarkably influence the annotation outcome and, therefore, the disagreement levels. Such individual differences can be partially explained by cultural and socio-demographic norms and variables, such as age, gender, instruction level, or cultural background. However, none of them is sufficient to capture the uniqueness of each subject and their evaluations.

Stimulus Characteristics. Instance characteristics have paramount importance for the annotation as well. Language meaning is often equivocal and carries ambiguities of several kinds: lexical, syntactical, semantic, and others. Humour, for example, often relies on lexical or syntactic ambiguity (Raskin, 1985; Poesio, 2020). Other genres using deliberate ambiguity as a rhetorical device include poetry (Su, 1994) or political discourse (Winkler, 2015).

For some instances, more than one label is correct, and the relative annotation task would be better framed as multi-label multi-class, rather than as multi-class *tout-court*. This is a common scenario in image and text tagging, where several object/features/topics can be present: this layer of complexity is a further potential source of disagreement between coders.

Context. Last but not least, the context matters. The same coder could give different answers at different times to the same questions. The answers change as the subjects' state of mind does, and even factors such as attention slips play a non-negligible role (Beigman Klebanov et al., 2008). This lack of consistency in human behavior is well known and explored in longitudinal studies, not only in psychology but also in linguistics (Lin and Chen, 2020).

These three aspects suggest that squeezing the human experience and resulting annotation into a set of crisp variables is a gross oversimplification in most cases.

2.2 Disagreement in 'Objective' Tasks

The NLP community has long been aware that it makes no sense to evaluate natural language generation applications against a hypothetical 'gold' output. These areas have developed specialized training and evaluation methods (Papineni et al., 2002; Lin, 2004). More surprisingly, disagreements in interpretation have been found to be frequent in annotation projects concerned with apparently more 'objective' aspects of language, such as coreference (Poesio and Artstein, 2005; Recasens et al., 2011), part-of-speech tagging (Plank et al., 2014), word sense disambiguation (Passonneau et al., 2012) and semantic role labelling (Dumitrache et al., 2019), to name a few examples. Even if in these tasks individual instances can be found to be reasonably objective, these findings appear to reflect the existence of extensive and systematic disagreement on what can be concluded from a natural language statement (Pavlick and Kwiatkowski, 2019).

2.3 Disagreement on 'Subjective' Tasks

Disagreement in annotation has been studied from a particular angle when occurring in highly subjective tasks such as offensive and abusive language detection or hate speech detection. Akhtar et al. (2019) introduced the *polarization index*, aiming at measuring a particular form of disagreement stemming from clusters of annotators whose opinions on the subjective phenomenon are polarized, e.g., because of different cultural backgrounds. Specifically, polarization measures the ratio between intra-group and inter-group agreement at the individual instance level, capturing the cases where different groups of annotators strongly agree on different labels. In this view, polarization is a somewhat complementary concept to disagreement, whereas a set of annotations could exhibit the latter but not the former, or both. Akhtar et al. (2020) employs this polarization measure to extract alternative gold standards from a dataset annotated with hate speech and train multiple models in order to encode different perspectives on this highly subjective task. While it clearly appears that involving the victims of hate speech in the annotation process helps uncovering implicit manifestations of hatred, the study also shows that the plurality of perspectives is more informative than the mere sum of the annotations.

3 Evaluation in Light of Disagreement

While the research mentioned in the previous section questions the assumption that a single 'hard' label (a gold label) exists for every item in a dataset, the models proposed for learning from multiple interpretations are still largely evaluated under this assumption, using 'hard' measures like Accuracy or class-weighted F1 (Plank et al., 2014; Rodrigues and Pereira, 2018).

Abandoning the gold standard assumption requires the ability to evaluate a system's output also over instances on which annotators disagree. There is no consensus yet on this form of evaluation, but a few proposals have been used already.

In fact, a way of performing soft evaluation exists which is a natural extension of current practice in NLP. This is to evaluate ambiguity-aware models by treating the probability distribution of labels they produce as a **soft label**, and comparing that to a full distribution of labels, instead of a 'one-hot' approach. This can be done using, for example, cross-entropy, although other options also exist. This approach was adopted in, *inter alia*, (Peterson et al., 2019; Uma et al., 2020; Fornaciari et al., 2021). Peterson et al. (2019) tested this approach on image classification tasks, generating the soft label by transforming the item annotation distribution using standard normalization. Uma et al. (2020) employed this form of soft metric evaluation for NLP, also comparing different ways to obtain a soft label from the raw data. They use soft metrics to compare the classifiers' distribution to the human-derived label distributions, complementing traditional hard evaluation measures.

Basile (2020) suggested a more extreme evaluation framework, where a model is required to produce different outputs encoding the individual annotators' labels. The predictions are then individually evaluated against the single annotations, rather than against an aggregated gold standard. This proposal aims at fostering the design of 'inclusive' models with respect to diverse backgrounds in highly subjective tasks.

While evaluating with disagreement is not yet widely adopted, methods for doing so exist. In the rest of this section, we discuss the two aforementioned approaches more in detail.

3.1 The SEMEVAL 2021 Campaign

The objective of SEMEVAL-2021 Task 12 on Learning with Disagreements (LeWiDi) (Uma et al., 2021) was to provide a unified testing framework for learning from disagreements in NLP and CV using datasets containing information about disagreements for interpreting language and classifying images.

Five well-known datasets for very different NLP and CV tasks were identified, all characterized by a multiplicity of labels for each instance, by having a size sufficient to train state-of-the-art models, and by evincing different characteristics in terms of the crowd annotators and data collection procedure. These include: a dataset of Twitter posts annotated with POS tags collected by Gimpel et al. (2011), a datasets for humour identification by Simpson et al. (2019), and two CV datasets on object identification namely the LabelMe (Russell et al., 2008) and CIFAR-10 datasets (Peterson et al., 2019).

Both hard evaluation metrics (F1) and soft evaluation metrics (cross-entropy, as discussed in Section 3) were used for evaluation (Uma et al., 2021). The results showed that in nearly all cases, models that account for noise and disagreement have the best (lowest) cross-entropy scores. These results are consistent with the findings of Uma et al. (2020) and Peterson et al. (2019).

3.2 Evaluation of Highly Subjective Tasks

Basile (2020) explored the impact of disagreement caused by polarization on evaluation, focusing on NLP tasks with high levels of subjectivity. They argue that aggregated test sets lead to unfair evaluation concerning the multiple perspectives stemming from the annotator's background. Therefore, they argue for a paradigm shift in NLP evaluation, where benchmarks for highly subjective tasks should consider the diverging opinions of the annotators throughout the entire evaluation pipeline.

This proposal is tested with a simulation on synthetic data, where the annotation is conditioned on two input parameters: difficulty (as in general ambiguity of the annotation task) and subjectivity (an annotation bias linked to a predetermined background variable for the annotators). They propose a straightforward evaluation framework that accounts for multiple perspectives on highly subjective phenomena, where multiple models are trained on the annotations provided by individual annotators, and their accuracy is averaged as a final evaluation metric. The findings from the experiment show that subjectivity and ambiguity are discernible signals, as discussed in Section 2. Moreover, it is shown how a perspective-aware framework provides a more stable evaluation for classifiers of highly subjective tasks, very much in line with the results by Uma et al. (2020).

4 Conclusion

In this position paper, we argue against the current prevalent evaluation practice of comparing against a single truth. This method has allowed automated evaluation, sped up model selection and development, and resulted in good evaluation scores. However, those scores hide the truth about the state of our models: many tasks are complex and subjective. Assuming a single truth for the sake of evaluation amounts to a gross oversimplification of inherently complex matters. We further reject the notion that we should remove annotation noise from datasets. Instead, we propose to embrace the complex and subjective nature of task labels. We show how disagreement from the annotator, the data, and the context, affects even seemingly objective tasks. Research already shows that incorporating this disagreement leads to better training performance. We suggest that it can do the same for evaluation. The datasets already exist, all we need is to use them. It might not produce the same nice high scores we have gotten used to. But it will provide an honest assessment of how good our models are, and do justice to the complexity of the subject we are trying to model.

Acknowledgements

This research was supported in part by the DALI project, ERC Grant 695662, and the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR) and by the Independent Research Fund Denmark (DFF) grants No. 9131-00019B and 9063-00077B. TF and DH are members of the MilaNLP group, and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AI*IA 2019 – Advances in Artificial Intelligence*, pages 588–603, Cham. Springer International Publishing.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):151–154.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *Proc. of the AIXIA Workshop*. Universitá di Torino.

Beata Beigman Klebanov and Eyal Beigman. 2009. Squibs: From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing disagreements. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 2–7, Manchester, UK. Coling 2008 Organizing Committee.

Samuel R Bowman and George E Dahl. 2021. What will it take to fix benchmarking in natural language understanding? *arXiv preprint arXiv:2104.02145*.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. SemEval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 100–105, Boulder, Colorado. Association for Computational Linguistics.

Kenneth Ward Church and Joel Hestness. 2019. A survey of 25 years of evaluation. *Natural Language Engineering*, 25(6):753–767.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2164–2170, Minneapolis, Minnesota. Association for Computational Linguistics.

Tommaso Fornaciari, Silviu Uma, Alexandra Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, Oregon, USA. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, USA. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Nancy Ide and James Pustejovsky, editors. 2017. *The Handbook of Linguistic Annotation*. Springer.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297, Lisbon, Portugal. Association for Computational Linguistics.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562, Atlanta, Georgia. Association for Computational Linguistics.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

You-Min Lin and Michelle Y. Chen. 2020. Understanding writing quality change: A longitudinal study of repeaters of a high-stakes standardized english proficiency test. *Language Testing*, 37(4):523–549.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252.

Rebecca J. Passonneau and Bob Carpenter. 2013. The benefits of a model of annotation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 187–195, Sofia, Bulgaria. Association for Computational Linguistics.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Massimo Poesio. 2020. Ambiguity. In Daniel Gutzmann, Lisa Matthewson, and Cécile Meier and Hotze Rullmann and Thomas Ede Zimmermann, editors, *The Companion to Semantics*. Wiley.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.

Victor Raskin. 1985. *Semantic Mechanisms of Humor*. D. Reidel, Dordrecht and Boston.

Marta Recasens, Ed Hovy, and M. Antonia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Philip Resnik and Jimmy Lin. 2010. 11 evaluation of nlp systems. *The handbook of computational linguistics and natural language processing*, 57.

Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. In *AAAI Conference on Artificial Intelligence*.

Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. LabelMe: A database and Web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728, Florence, Italy. Association for Computational Linguistics.

Soon P. Su. 1994. *Lexical Ambiguity in Poetry*. Longman, London.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, and Massimo Poesio. 2021. Semeval-2021 task 12: Learning with disagreements. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft-loss functions. In *Proceedings of the 8th AAAI Conference on Human Computation and Crowdsourcing*, pages 173–177.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Susanne Winkler, editor. 2015. *Ambiguity: Language and Communication*. De Gruyter.